Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics

Xavier Roca,1 Andrew J. Olson,1 Atmakuri R. Rao,1,6 Espen Enerly,2,3 Vessela N. Kristensen,2,3 Anne-Lise Børresen-Dale,2,3 Brage S. Andresen,4,5 Adrian R. Krainer,1 and Ravì Sachidanandam1,7

1Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; 2Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Centre, Montebello 0310, Oslo, Norway; 3Faculty of Medicine, University of Oslo, Norway; 4Department of Human Genetics, Aarhus University, Aarhus 8000C, Denmark; 5Aarhus University Hospital, Sygehus 8000N, Denmark

Many human diseases, including Fanconi anemia, hemophilia B, neurofibromatosis, and phenylketonuria, can be caused by 5′-splice-site (5′ss) mutations that are not predicted to disrupt splicing, according to position weight matrices. By using comparative genomics, we identify pairwise dependencies between 5′ss nucleotides as a conserved feature of the entire set of 5′ss. These dependencies are also conserved in human–mouse pairs of orthologous 5′ss. Many disease-associated 5′ss mutations disrupt these dependencies, as can some human SNPs that appear to alter splicing. The consistency of the evidence signifies the relevance of this approach and suggests that 5′ss SNPs play a role in complex diseases.

[Supplemental material is available online at www.genome.org.]

The sequenced genomes of a wide range of organisms allow global, comparative analyses of regulatory sequences. The genomic set of splice-site sequences corresponds to a large-scale splicing experiment performed by nature under evolutionary constraints. Here we focus on 5′-splice-site (5′ss) sequences of the U2-type GT-AG class, which comprise over 98% of all splice sites, and use disease-causing mutations, human single nucleotide polymorphisms (SNPs), and variations in natural splice sites in the genome (within and between species) to infer properties inherent to 5′ss, with important implications for human genetics.

Splice sites are conserved sequences at both ends of an intron that are recognized during the initial steps of splicing (Hastings and Krainer 2001; Brow 2002; Jurica and Moore 2003). Both the 5′ss and the 3′ splice site (3′ss) conform to degenerate motifs that are recognized by specific splicing factors. The U2-type GT-AG 5′ss, spanning 3 nucleotides (nt) at the 3′ end of the exon and 6 nt at the 5′ end of the intron, is initially recognized via base pairing to the 5′ end of the U1 snRNA (Fig. 1; Zhuang and Weiner 1986; Séraphin et al. 1988; Siliciano and Guthrie 1988). Later in the reaction, U5 and U6 snRNAs base pair to exonic or intronic 5′ss positions, respectively (Newman and Norman 1992; Wassarman and Steitz 1992; Kandels-Lewis and Séraphin 1993; Lesser and Guthrie 1993; Crotti et al. 2007). Additional elements influence splice-site selection, such as exonic or intronic splicing enhancers (ESE, ISE) or silencers (ESS, ISS) (Cartegni et al. 2002; Ladd and Cooper 2002).

Even though the mammalian 5′ss consensus sequence (CAG/(GTAAGT) is perfectly complementary to the 5′ end of U1 snRNA, individual 5′ss exhibit considerable variation at different positions, indicating a tolerance for mismatches in U1 base pairing. The free energy of the 5′ss/U1 base pairing is not always a good predictor of 5′ss efficiency (Roca et al. 2005), suggesting the existence of other factors that influence 5′ss selection. It is known that proteins such as the U1-C polypeptide (Du and Rosbash 2002), which is a component of the U1 snRNP, PRPF8 (also known as PRP8) (Maroney et al. 2000), and members of the SR and hnRNP protein families (Mayeda and Krainer 1992; Cáceres et al. 1994; Buratti et al. 2004) are also involved in 5′ss selection. Most methods for estimating 5′ss efficiency are based on position weight matrices (PWMs) that are calculated from collections of splice sites (for a depiction of a PWM, see Methods and Fig. 1) (Shapiro and Senapathy 1987; Senapathy et al. 1990). Compared with PWMs, which assume independence between positions within the 5′ss, methods that have considered the dependencies between these positions (Brunak et al. 1991; Yeo and Burge 2004) have provided some improvements in predicting 5′ss efficiency (Roca et al. 2003, 2005; Buratti et al. 2007). Such analyses have been previously carried out on smaller data sets (Burge and Karlin 1997; Carmel et al. 2004) or implicitly considered in various 5′ss scoring algorithms (Brunak et al. 1991; Yeo and Burge 2004; Krawczak et al. 2007).

Mutations at 5′ss are frequent among mutations that cause human disease, from genetic disorders to cancer (Krawczak et al. 1992; Nakai and Sakamoto 1994; Teraoka et al. 1999; Stenson et al. 2003; Buratti et al. 2007). In many cases, the mutations cause aberrant splicing by affecting the invariant GT dinucleotide at positions +1 and +2 of the 5′ss. However, the effects of other disease-causing 5′ss mutations are less clear-cut: Nucleotide substitutions affecting the less conserved positions can cause splicing defects in some but not all 5′ss, suggesting that the remaining 5′ss positions and/or the overall context dictate the extent to which splicing is disrupted.

We present a comprehensive analysis of the pairwise associations between nucleotides at different 5′ss positions, using the
splice-site compilation from SpliceRack (Sheth et al. 2006), which is a collection of more than 500,000 splice sites from five genomes. We explicitly and exhaustively identify combinations that appear to be significant. We show that there are constraints on combinations of nucleotides at different positions of the 5’ss. We further show that these constraints are indeed operational in the evolutionary process using four data sets: disease-causing mutations at 5’ss, orthologous pairs of 5’ss between mouse and human, SNPs at 5’ss, and simulated SNPs. It is remarkable that all these disparate sources of information seem to be consistent with a single explanation: that the pairwise associations between positions of the 5’ss are a determinant of 5’ss efficiency.

Results

The primary determinant of the efficiency of a 5’ss is its match to the PWM (Zhuang and Weiner 1986; Séraphin et al. 1988; Sili- ciano and Guthrie 1988) given by the log-odds score (Methods), which reflects, at least in part, the stability of base pairing to the 5’s end (Fig. 1). However, many 5’ss with similar PWM scores show differences in splicing efficiency, suggesting secondary effects that are ignored by the PWM. To study additional determinants of 5’ss efficiency, we have analyzed the associations between pairs of nucleotides at different positions of the U2-type GT-AG 5’ss in five species: Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans, and Arabidopsis thaliana.

Two-nucleotide associations

PWMs for each of the five species can be calculated from the genomic 5’ss data in SpliceRack (Sheth et al. 2006), consisting of 183,678 H. sapiens 5’ss, 174,671 M. musculus 5’ss, 40,367 D. melanogaster 5’ss, 93,699 C. elegans 5’ss, and 111,351 A. thaliana 5’ss. The PWMs are calculated by counting the occurrences of nucleotides at each position within the 5’ss. The expected occurrence of pairs of nucleotides at two positions, e.g., A at +4 and G at +5, can be calculated from the PWM frequencies by assuming independence between the two positions (for nomenclature of positions in 5’ss, see Fig. 1). The actual frequency of occurrence for such pairs can be measured from the genomic data, by counting the occurrence of pairs of nucleotides (16 possible combinations) at all possible combinations of positions (21 unique combinations, if the +1 and +2 invariant positions are excluded). The deviations of the actual count from the expected count are measured using a log-odds scoring scheme (see Methods), and for visualization, the results are scaled and shown on a colored matrix. A combined matrix showing data from M. musculus and H. sapiens is depicted in Figure 2; a matrix for C. elegans and A. thaliana is shown in Supplemental Figure S1; and the D. melanogaster matrix is shown in Supplemental Figure S2.

The surprising observation is that the pattern of deviations of the actual counts from the expected counts is, to a great extent, conserved between species. This suggests that these patterns of deviation are a result of the mechanisms of 5’ss recognition by the splicing machinery (e.g., cf. upper and lower triangles in Fig. 2 as well as Supplemental Figs. S1, S2). The mouse and human patterns show remarkably close similarity. We show the variability for a few pairs within the human genome and across genomes in Table 1. There are some differences, but the overall patterns of deviation are clearly preserved and presumably reflect the pressures that arise from the conservation of the splicing machinery.

The depletion (maroon coloration) in the parts of the matrix that connect the exonic and intronic portions of the 5’ss is a striking feature in all species (Figs. 2, S1, S2; see associations between positions −2 or −1 and +3, +4, +5, or +6). This implies that having a nonconsensus nucleotide on the exonic side causes a depletion of nonconsensus nucleotides on the intronic side, and vice versa, which is consistent with the proposal of a seesaw linkage pattern (Burge and Karlin 1997; Carmel et al. 2004). Some combinations are surprising in the distance they span, such as the enhancement of the pair −1T+6T. The pair −1C+3C shows a strong depletion in all species, for reasons unknown to us. Another striking pattern is the association between positions −1 and +5 wherein a G (consensus) at −1 allows any nucleotide at position +5; similarly, +5G (consensus) allows any nucleotide at −1. We show some of the conserved pairwise associations in Table 1.

Some features in the pairwise association matrix might be the result of constraints other than the splicing mechanism. For example, the combinations −3T−2A and −2T−1A are severely depleted in the five species, probably because these combinations can be part of two of the three stop codons (TAG and TAA). The combination +5C+6C is enriched, probably reflecting the gradual conversion of U12-type GT-AG introns into U2-type GT-AG introns (Burge et al. 1998; Sheth et al. 2006). From the pairwise association matrix, it can be seen that the CpG dinucleotide is depleted in higher species, whereas D. melanogaster and C. elegans do not show such depletion, consistent with the fact that in higher organisms CpG is underrepresented, due to methylation of the C followed by de-amination (Tweedie et al. 1997). The spliceosome could have evolved to use patterns that might be species-specific, even though the drive to maintain the patterns might come from other processes.
Another possible explanation for some of the pairwise associations could lie in nucleotide biases in error-prone DNA repair. However, the dependence of these mechanisms on the neighboring nucleotides is not very pronounced (Krawczak et al. 1998), arguing against a large contribution to the associations.

There are some species-specific features in the pairwise associations. For example, there are tighter associations (brighter blues) between the exonic nucleotides at positions –2 and –1 in mammals and A. thaliana than in the two invertebrates (Figs. 2, S1, S2). In contrast, D. melanogaster and C. elegans have more biases between intronic nucleotides (positions +3 to +6 show brighter blues).

The human genome is organized into isochores, i.e., regions with relatively stable GC content compared with the variation of the GC content across the genome (Costantini et al. 2006). The isochores have been classified into five groups (H3 [GC > 53.5%], H2 [46.6–53.5%], H1 [41.6–46.5%], L2 [37.6–41.5%], and L1 [<37.5%]). In order to use this stratification to study variations in the associations, the 5’ss were classified into the five isochore categories, and the expected and actual counts for pairs within each set were calculated. In Table 1 we show the variation of the ratios of actual to expected counts for a selection of pairs. Overall, the associations are preserved, both within and across species.

Implications for the splicing machinery

A striking feature that is conserved across species is the depletion of U1-noncomplementary pairs across the exon–intron bound-
Table 1. A selection of pairwise associations in 5’ss and their variations within and across genomes

<table>
<thead>
<tr>
<th>Pair</th>
<th>Actual/expected (H)</th>
<th>Ratio (C.I.) (H)</th>
<th>Ratio (isochores)</th>
<th>Ratio (C,D,A,M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>–3C –2G</td>
<td>3637/7758</td>
<td>0.47 (0.46–0.48)</td>
<td>0.341–0.578</td>
<td>0.714,0.813,0.700,0.437</td>
</tr>
<tr>
<td>–2C –1G</td>
<td>9744/15,998</td>
<td>0.61 (0.60–0.62)</td>
<td>0.403–0.779</td>
<td>0.935,0.921,0.743,0.592</td>
</tr>
<tr>
<td>+2T –1A</td>
<td>1095/2598</td>
<td>0.42 (0.405–0.443)</td>
<td>0.334–0.448</td>
<td>0.434,0.376,0.419,0.401</td>
</tr>
<tr>
<td>–1C+3C</td>
<td>12/129</td>
<td>0.09 (0.051–0.135)</td>
<td>0.002–0.020</td>
<td>0.569,0.139,0.099,0.082</td>
</tr>
<tr>
<td>–1C–5T</td>
<td>9/371</td>
<td>0.024 (0.001–0.04)</td>
<td>0.000–0.008</td>
<td>0.391,0.234,0.266,0.038</td>
</tr>
<tr>
<td>+1G–5A</td>
<td>15,887/15,233</td>
<td>1.20 (1.196–1.202)</td>
<td>1.158–1.265</td>
<td>1.375,1.285,1.166,1.194</td>
</tr>
<tr>
<td>+3T–4T</td>
<td>1484/538</td>
<td>2.758 (2.684–2.872)</td>
<td>1.544–2.365</td>
<td>2.099,2.440,1.147,2.499</td>
</tr>
<tr>
<td>+3T–5T</td>
<td>129/352</td>
<td>0.366 (0.326–0.429)</td>
<td>0.012–0.134</td>
<td>0.876,0.309,0.894,0.276</td>
</tr>
<tr>
<td>+3C–4T</td>
<td>816/556</td>
<td>1.467 (1.407–1.561)</td>
<td>1.036–2.844</td>
<td>0.384,1.1,436,1.679</td>
</tr>
<tr>
<td>+3C–5T</td>
<td>111/371</td>
<td>0.299 (0.267–0.364)</td>
<td>0.005–0.255</td>
<td>0.327,0.120,0.819,0.261</td>
</tr>
<tr>
<td>+3A–5G</td>
<td>74/445</td>
<td>0.166 (0.153–0.219)</td>
<td>0.041–0.204</td>
<td>0.159,0.162,0.635,0.181</td>
</tr>
<tr>
<td>+3A+5G</td>
<td>75,891/86,526</td>
<td>0.877 (0.875–0.878)</td>
<td>0.858–0.933</td>
<td>0.896,0.932,0.883,0.892</td>
</tr>
<tr>
<td>+3G+5A</td>
<td>1856/5753</td>
<td>0.322 (0.311–0.333)</td>
<td>0.273–0.385</td>
<td>0.605,0.379,0.676,0.319</td>
</tr>
<tr>
<td>+4C–5C</td>
<td>2431/705</td>
<td>3.448 (3.275–3.460)</td>
<td>1.443–2.246</td>
<td>0.666,0.354,1.094,3.326</td>
</tr>
<tr>
<td>+5C+6G</td>
<td>1206/1948</td>
<td>0.619 (0.595–0.649)</td>
<td>0.485–0.883</td>
<td>0.940,1.117,0.679,0.616</td>
</tr>
</tbody>
</table>

The confidence interval (C.I.) was established by using bootstrapping. The second and third columns show data for human splice sites, and the ratios refer to actual count/expected count. We also used the five types of isochores in the human genome (H) (Costantini et al. 2006) to establish the variation in the ratio of actual to expected counts (since random sampling may hide the stratification in the data). The isochores have an effect on adjacent pairs that are CpG and on pairs that do not occur often, but the bias is always in the direction seen with the full genomic data set, confirming that the depletion or elevation of pairs is under some selection pressure. The isochore numbers and the numbers for the species (A. thaliana [A], M. musculus [M]) other than H. sapiens (H) correspond to the full data set, and the confidence intervals were not determined. Not all splice sites map to isochores, and in addition, only sites that map uniquely were used in this calculation. They are shown to indicate the variability of the ratios across the data sets.

We observed that the pair +3C–4T is enhanced in the five species. This association likely reflects base pairing to the U6 snRNA. All of the associations are part of the U5/C snRNAs (Hiser and Guthrie 1993). The remaining nucleotides that would base pair to U6 but not to U1 do not show an association that is phylogenetically conserved.

There are a number of biased combinations that cannot be explained by base pairing to any of these snRNAs, such as an enhancement of +4C–5C in all species, or –2C –1T in the vertebrates. Most importantly, these tend to be less conserved than the previous associations. It is possible that these combinations are part of binding sites for proteins that influence 5’ss selection, such as U1-C, PRPF8, SR proteins, or hnRNP proteins (Mayeda and Krainer 1992; Cáceres et al. 1994; Maroney et al. 2000; Du and Rosbash 2002; Buratti et al. 2004). Whether these dinucleotide associations reflect a portion of a protein binding site and whether the binding specificity of some of these proteins is slightly different between species remain to be elucidated.

Scoring the pairwise associations within 5’ss

We used the pairwise association matrix and a log-odds scoring scheme to score 5’ss for their pre-
Determinants of 5′ss efficiency

Figure 3. Scatter plots showing combinations of PWM scores (X-axis) and association-score differences (Y-axis) for 5′ss from four data sets. The four data sets are disease-causing mutations (A, 595 cases), orthologous mouse-human pairs of 5′ss (B, 19,940 pairs), SNPs at 5′ss (C, 1,260 SNPs), and simulated SNPs at 5′ss (D, 8074 cases). Each spot in the data set corresponds to a 5′ss and a single-nucleotide change to the 5′ss. In each data set, along the X-axis, we plotted the original 5′ss PWM score; in the case of SNPs and orthologous 5′ss, the 5′ss with the higher score is considered the original 5′ss. Along the Y-axis, we plotted the difference in association scores due to the change (score of original pair – score of new pair). A change to a pair with reduced association score will give a positive score difference, and an increase in association score will give a negative value. The solid lines show the averages along the X- and Y-axes, whereas the dashed lines show the confidence intervals for the means (calculated by a bootstrap method). The numbers are also given as text within the plots, with the confidence interval shown in parentheses. The disease panel is clearly biased toward low PWM scores and greater disruption of associations, compared with the other data sets. The averages for the four data sets are well separated on the basis of the confidence intervals.

(Reedy et al. 1981) in the 5′ arm of the U1 snRNA (Fig. 1). Thus, it is not obvious why the A to G mutations at +3 in genes such as ACADSB (also named SBCAD), BRCA1, APC, and NF1 can cause genetic diseases (Madsen et al. 2006). It has been shown that 5′ss with disease-causing +3 A-to-G mutations exhibit two distinct features: (1) They are intrinsically weak (Madsen et al. 2006); and (2) they frequently have nonconsensus nucleotides at positions +4 and +5 (for confirmation, see Fig. 4; Ohno et al. 1999; Madsen et al. 2006). Figure 4 also shows that 5′ss with G at position +3 and SNPs with A/G alleles at position +3 prefer consensus nucleotides at +4 and +5, in agreement with the second observation.

These observations can be explained by the dependences between position +3 (A/G) and the nucleotides at positions +4 and +5 (Fig. 2). The association of +3A to nonconsensus nucleotides at +4 (C, G, T) and +5 (A, C, T) is blue (enhanced), whereas the association of +3G to nonconsensus nucleotides at +4 and +5 is maroon (depleted). An interesting prediction is that if both +4 and +5 are nonconsensus, then the splicing defect in the mutant can be fixed by converting either +4 or +5 independently to the consensus (see below).

Experimental tests of the associations

We carried out in vitro splicing experiments in order to confirm our prediction for the +3 A-to-G mutations. We used a 5′ss-competition assay described in Figure 5. We inserted variants of the 5′ss (GGG/GUACAU) from exon 3 of the ACADSB gene into a beta-globin (HBB) minigene, which allows direct comparison of the efficiencies of any two 5′ss (Roca et al. 2005). An A-to-G mutation at position +3 of this 5′ss causes a rare metabolic disease, ACADSB deficiency (Madsen et al. 2006).

First, we found that the +3 A-to-G mutation in the ACADSB 5′ss severely reduces the 5′ss strength (Fig. 6). When the ACADSB wild-type 5′ss (+3A) or its +3 A-to-G mutant version (+3G) were tested in competition against the beta-globin cryptic 5′ss at −16 (Treisman et al. 1983), only the cryptic 5′ss was used (lanes 1, 2). Since the cryptic 5′ss is already a suboptimal 5′ss (Roca et al. 2003), we conclude that both the +3A and +3G 5′ss are weak. When the +3A and +3G 5′ss were competing against each other, splicing occurred only via the +3A 5′ss (lanes 4, 5). This finding indicates that the +3A 5′ss is much stronger than its +3G counterpart.

Second, we found that correcting positions +4 and/or +5 to match the consensus can alleviate the effects of the +3 A-to-G mutation, which is in concordance with our above-mentioned prediction derived from the pairwise associations. To test for the rescue of splicing by correcting these positions, we compared two ACADSB 5′ss with the same combination of nucleotides at positions +4, +5, and +6, but one having A at +3 and the other one G at +3 (Fig. 7). In general, we found that +3G use was positively correlated with the number of consensus nucleotides at positions +4 to +6 (lanes 1–3, 5). However, the two reciprocal
Non-consensus Nucleotides

![Percentage of nonconsensus nucleotides at positions (X-axis) of the 5'ss for five classes of 5'ss: all 5'ss in human (186,630 cases), 5'ss with A at +3 (110,598), 5'ss with G at +3 (64,837), 5'ss with SNPs having alleles A/G at position +3 (93), and 5'ss with disease-causing A-to-G mutations at +3 (50 cases). 5'ss with disease-associated A-to-G mutations at position +3 tend to have nonconsensus nucleotides at positions +4 and +5 more frequently than the other sets, which suggests their disruptive nature, based on the association matrix (Fig. 2).](image)

Figure 4. Percentage of nonconsensus nucleotides at positions (X-axis) of the 5'ss for five classes of 5'ss: all 5'ss in human (186,630 cases), 5'ss with A at +3 (110,598), 5'ss with G at +3 (64,837), 5'ss with SNPs having alleles A/G at position +3 (93), and 5'ss with disease-causing A-to-G mutations at +3 (50 cases). 5'ss with disease-associated A-to-G mutations at position +3 tend to have nonconsensus nucleotides at positions +4 and +5 more frequently than the other sets, which suggests their disruptive nature, based on the association matrix (Fig. 2).

experiments shown in this figure do not match, in the sense that the activation of the +3G 5'ss is influenced by the relative position of the two competing 5'ss (lanes 1–6, cf. the +3G-spliced band between both panels). This difference is probably due to the influence of the sequences that flank the 5'ss at positions −16 and +1. Notwithstanding these positional effects, our conclusion that correction of positions +4 and/or +5 rescues splicing of the +3G 5'ss is still supported by the data.

Finally, the +3G 5'ss can be activated by correcting position +4 alone (Fig. 7, lane 3). This is not surprising, because a base pair at +4 with U1 might stabilize the wobble G·Ψ base pair at +3. Strikingly, correcting position +5 alone alleviated the severity of the A-to-G mutation at +3 (Fig. 7, lane 5). This indicates that nonadjacent interactions (such as those involving positions +3 and +5) at the 5'ss can affect splice-site efficiency.

Orthologous 5'ss between the mouse and human genomes

We expected that orthologous pairs of 5'ss between mouse and human genomes, which are very similar to each other according to the PWMs and the association matrices, should show patterns of nucleotide substitutions that do not substantially disrupt the PWM score or the association score. We used a published data set of 54,761 pairs of orthologous 5'ss from the human and mouse genomes (Carmel et al. 2004) to study the nature of differences between members of these pairs.

We found that 36.5% of these pairs have identical splice-site sequences and 36.4% show a single nucleotide change. The nucleotide change between orthologous pairs that differ at a single position does not usually disrupt the PWM score and the pairwise associations within 5'ss (Fig. 3). We also observed a compensatory trend: If one member of an orthologous 5'ss pair has a weaker PWM score, then its pairwise-association score tends to be higher (Fig. 8).

In the orthologous 5'ss pairs with two nucleotide changes (19.2%), certain coordinated changes are preferred over others. For example, +5C+6G changes to +5T+6G more often than expected (72 vs. 28 expected; both combinations are strongly enriched), whereas −1T+6T to −1C+6A is strongly suppressed (occurs once vs. 20 expected; the first combination is enhanced, whereas the second combination is strongly depleted). The probability of a single change occurring, e.g., +5C to +5T, is calculated by counting the number of such transitions over the pairs of which at least one of them has a C at +5. The expected number of coordinated pairwise changes is calculated by assuming independence between the transitions at the two positions in the 5'ss. Evidently, the pairwise-association patterns reflect constraints on the evolution of 5'ss due to the conserved splicing machinery, and these association patterns could hold clues into mechanistic features of the splicing machinery.

SNPs at 5'ss

We expect SNPs that occur at splice sites to be neutral, i.e., not to affect splice-site efficiency in a majority of the cases. There have been a few reports of SNPs that do affect splicing (Shimamura et al. 2004; Field et al. 2005; Skarratt et al. 2005; Elsharaway et al. 2006; Kawase et al. 2007) but no comprehensive analysis of this phenomenon. In our study, we used the 1260 SNPs at 5'ss (dbSNP) (Smigielski et al. 2000) that are confirmed by population data and lie outside of positions +1 and +2.

We calculated the PWM scores and pairwise-association score differences for SNPs (Fig. 3C). The average PWM score for 5'ss with SNPs (5.92) is higher than that for simulated SNPs (5.54), which in turn is higher than the average score for disease-causing mutations (5.31). The PWM scores of orthologs that show one difference between human and mouse are higher on average than the SNP set (6.2 vs. 5.92), suggesting that strong 5'ss (PWM score >6.2) can tolerate such changes without affecting splicing.
From Figure 3, we can see that the disruption of the associations by the SNPs is smaller than the disruption by the simulated SNPs, which in turn is smaller than the disruption by the 5's disease-causing mutations, suggesting selective pressure on SNPs to preserve the associations. For instance, Figure 4 shows that SNPs with alleles A and G at position +3 tend to have consensus nucleotides at positions +4 and +5, as expected from the previous discussion on 3-A-to-G disease mutations.

By using these findings (see Fig. 3), we were able to establish criteria to identify SNPs that might disrupt splicing. We used the average score of the disease set to identify low PWM scores (<3.31). The separation between disease-causing mutations and the SNPs and orthologous sets was used to identify high PWM-score differences (>1). The average association-score difference for the disease data set was used to identify high association-score differences (>1.1). We expect disruptive 5's SNPs to occur at 5's with low PWM scores, and for the SNPs to substantially disrupt either the PWM score or the association scores or both simultaneously.

**Disruptive 5's SNP predictions and evidence for their effect on splicing**

By using the criteria listed above, we created five sets from the 1260 SNPs in dbSNP that occur at 5's (Smigielski et al. 2000). The sets were picked to dissect the effects of the PWM scores, PWM-score differences between SNP alleles, and association-score differences between SNP alleles. The first set, designated LHH, consists of 35 SNPs at 5's that have low (L) PWM scores and alleles with high (H) association-score differences and high (H) PWM-score differences. The second set, LHL, consists of 35 SNPs at 5's with low PWM scores and alleles with high PWM-score differences and low association-score differences. The third set, LLH, consists of 43 SNPs at 5's with low PWM scores and alleles with low PWM-score differences and high-association-score differences. The fourth set, HLL, consists of 58 SNPs at 5's with high PWM scores and alleles with low PWM-score differences and low association-score differences. Finally, the fifth set, LLL, consists of 111 SNPs at 5's with low PWM scores and alleles with low PWM-score differences and low association-score differences. We used EST and mRNA alignments to the genome to establish disruption caused by SNPs at 5's, using the UCSC browser (Kent et al. 2002). The alignments were studied to pick unambiguous cases of intron retention (at least one intron besides the retained one was spanned by the EST/mRNA), exon skipping (the skipped exon is otherwise constitutively spliced), and cryptic splice-site utilization. A prediction was considered validated if alternative splicing was observed at the 5's in question, and when possible, the EST and mRNA data supporting the alternative event corresponded to the nonconsensus allele of the SNP. In the case of exon skipping, it is not possible to ascertain the allele at the SNP position, but by considering the proportion of internal exons among RefSeq mRNAs that are skipped in this manner (3734/166,622), we were able to infer the likely error rate in making the determination of exon skipping as being due to the SNP (one in 45). We estimate similar error rates for intron retention (one in 160) and cryptic-splice-site utilization (one in 70) based on previous work on the relative prevalence of the different types of alternatively spliced exons (Kim et al. 2007).

In agreement with our expectations, we found that the sets LHH (10 disruptive SNPs out of 35), LHL (eight disruptive SNPs out of 35), and LLH (12 disruptive SNPs out of 43) are more disruptive than the control sets HLL (five disruptive SNPs out of 58). The deleterious effect of the +3 A-to-G mutation can be alleviated by correcting positions +4 and/or +5. In this competition assay, we created two sequences from the ACADSB exon 3 5’s: one with +3A and another with +3G (for the experimental strategy, see Fig. 5; for the ACADSB 5’s sequences, see Fig. 6). The same mutations were simultaneously introduced at positions +4, +5, and +6 at both 5’s, to test their compensatory effects vis-à-vis the +3 A-to-G mutations. The nucleotides at positions +4, +5, and +6 are listed above the lanes, and the consensus nucleotides are boxed. The left and right panels are the same, except that the positions of the +3A and +3G 5’s are swapped (indicated on the top diagram). With a few exceptions, the general trend is that the +3G 5’s was used to a greater extent as the number of consensus nucleotides at positions +4 to +6 was increased (lanes 1–3, 5). The degree of use of the +3G 5’s was not the same for identical mutants in the two panels (for details, see text), but despite these positional effects, the trend remains. The +3G 5’s can be activated by correcting position +4 alone (lane 3) or +5 alone (lane 5). The rescue by +5 alone is surprising and indicates that associations between nonadjacent positions can affect splice-site efficiency.
SNPs at 5′ss, association scores improve as PWM scores get weaker. The graph shows the average of the PWM-score differences (white bars) as well as association-score differences (black bars) for the orthologous 5′ss mouse–human pairs, organized by the number of nucleotide differences between the members of each pair (x-axis). A positive difference means a decrease in the score (and vice versa). The differences in PWM scores are always positive, as the 5′ss with the higher score within a pair is always considered first. Error bars, confidence intervals.

Figure 8. Association improvements can compensate for reductions in PWM scores. In orthologous mouse–human pairs of 5′ss, association scores improve as PWM scores get weaker. The graph shows the average of the PWM-score differences (white bars) as well as association-score differences (black bars) for the orthologous 5′ss mouse–human pairs, organized by the number of nucleotide differences between the members of each pair (x-axis). A positive difference means a decrease in the score (and vice versa). The differences in PWM scores are always positive, as the 5′ss with the higher score within a pair is always considered first. Error bars, confidence intervals.

SNPs at 5′ss (19 disruptive SNPs out 111). By using bootstrapping to estimate the variance of these numbers, we found that the LLL and LLH sets are significantly different from each other (P-value of 0.0003 using an unpaired t-test). The stronger disruption shown by the LLH set compared with the LLL set confirms the predictive nature of the association scores. A sampling of SNPs from these sets is given in Table 2.

This analysis is confounded by the fact that the ESTs might not have sampled the SNP alleles of interest, due to the genotype of the source. The intronic SNPs are also not likely to be sampled in the ESTs. In addition, some of the alternative spliced products might be degraded by NMD, due to the presence of a premature stop codon and thus would not be observed (Lejeune and Maquat 2005). In light of these limitations, it is encouraging that we still find such a high confirmation rate.

We expected the genotypes to show a preference for alleles of SNPs at 5′ss that do not disrupt splicing, but we did not detect a strong signal in the distribution of genotypes. The effect of SNPs at 5′ss might be alleviated by the sequence context. For example, in the ACDADM (also called MCAD) gene (Nielsen et al. 2007), one allele of a SNP creates an ESS, and a mutation that disrupts a nearby ESE causes exon skipping only in individuals with the ESS-creating allele. Even if splicing is affected, there might be no phenotypic effects due to the following: (1) a small reduction in the correct mRNA levels; (2) compensation in the genetic network; (3) the function of the disrupted gene is required only under certain conditions; or (4) the alternate product due to the disruption of the particular 5′ss is also functional. Nevertheless, such SNPs can offer a window into the investigation of the genetic network. This analysis will allow investigation of disruptive SNPs and their role in complex diseases.

Discussion

We have shown that the effects of disease-causing mutations are often a result of the disruption of conserved patterns in associa-

...tions between nucleotides at different positions within the 5′ss, and SNPs that disrupt these pairwise associations tend to affect splicing. A majority of SNPs respect the associations, as expected from SNP neutrality. In addition, we have shown that orthologous 5′ss mouse–human pairs show changes that likewise respect the associations, which suggests the existence of selective pressure to maintain them. A set of simulated SNPs is more disruptive of associations than neutral SNPs but is better than the disease-causing mutations in this respect. This is expected from the lack of selective pressure on the simulated SNPs.

The conservation of the associations is indicative of selective pressures reflecting functional features of the splicing machinery, and allows inferences to be made about the underlying mechanisms. The pairwise associations confirm many known effects, but also suggest new areas for exploration. We found pairwise associations probably related to U1 base pairing, such as specific patterns involving consensus and nonconsensus nucleotides across the exon–intron boundary. We found one association probably related to base pairing to U6, but none related to U5. The long-range associations are likely related to protein–RNA interactions, and further experiments should shed light on them. Similar studies could be carried out using 3′ss. However, the longer span of this motif (Sheth et al. 2006) and its more plastic organization suggest that the pairwise associations at 3′ss will not reveal as many biases as the associations at 5′ss.

By using the data from disease-causing mutations in 5′ss, differences between orthologous pairs of mouse and human 5′ss, and genomic data for five species, we were able to generate criteria for prediction of splicing-disruptive SNPs at 5′ss. We have shown that circumstantial evidence from ESTs provides support for these predictions and encourages further experiments to study their effects in vivo.

The disruptive SNPs may provide insights into genetic networks. If a proven disruptive SNP is in Hardy-Weinberg equilibrium, it suggests that the genetic network is immune to the change, and this can be a starting point for investigating the reasons for this robustness of the network. Alternatively, such SNPs might be implicated in complex diseases, wherein their effects are apparent only under certain genetic and environmental conditions. For example, there has been work on SNPs affecting the p53 pathway, such as a SNP in the MDM2 gene that alters a transcription-factor binding site and hence the levels of p53 (Bond and Levine 2007). This results in phenotypes ranging from an effect on fertility, to the onset of tumor formation, and the response to various therapies, such as estrogen replacement. Our study is a first step in investigating the effect of SNPs at 5′ss on a genomic scale. The consequences of the disruptions could be stochastic in nature, leading to effects that might not be readily discernible in in vitro systems but might nevertheless affect the in vivo phenotypes.

This study illustrates the power of the convergence of different data sets for obtaining insights into mechanisms of gene
expression and for understanding the neutral and disease-causing nucleotide changes found in human populations.

Methods

Scoring splice sites with PWM

P0Ms reflect the frequencies of the 4 nt (A, C, G, and T) at each position of the splice site. P0Ms can be used to score a site by converting the frequencies into a log-odds score (log of the ratio of the actual frequency and the expected frequency) (Sheth et al. 2006).

Scoring associations within splice sites

For the associations, we calculated the expected frequencies for each combination of a pair of positions and nucleotides, using the PWM and assuming independence of positions from each other. We then calculated a log-odds score for the combination, based on the actual numbers observed in the genomic data set. The total association score for a given 5’s is the sum of the log-odds score for each nucleotide pair at the 5’s. The change in association score measures the disruption (or enhancement) of the pairwise associations due to changes at the 5’s.

In order to avoid small statistical fluctuations affecting the scores (e.g., an actual occurrence of three versus an expected number of 45 would give the relevant pair a strong negative score), we excluded those pairs that are expected less than 50 times in the genome. This does not affect the scoring of most splice sites, as these pairs are by definition rare.

Collection of data sets

From the HGMD database (Stenson et al. 2003), we collected 1772 disease-causing mutations affecting U2-type GT-AG 5’s sequences. After excluding mutations affecting +1 and +2 of the 5’s, we were left with 582 mutations affecting the positions −3 to −1 and +3 to +6, which we used in our analysis. An additional 13 5’s mutations changing a +3 A to G at 5’s were added from a separate source (Madsen et al. 2006). We randomly selected 5’s from SpliceRack (Sheth et al. 2006) and mutated a randomly selected position in the 5’s to generate a list of simulated SNPs in 5’s. From dbSNP (Smigielski et al. 2000), we extracted SNPs that have population data and mapped them to the genome to identify SNPs at 5’s. For the orthologous set, we used a published data set (Carmel et al. 2004).

In vitro splicing

In the 5’s competition assay, each test ACADSB 5’s replaces the beta-globin 5’s at positions −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.

The beta-globin wild-type 5’s and the cryptic 5’s at position −16 were replaced by various mutant ACADSB 5’s by PCR mutagenesis. We inactivated a second cryptic 5’s in beta-globin and the cryptic 5’s at position −16 and/or +1 (Fig. 5). The competition between the two 5’s provides an estimate of the difference in their intrinsic strengths.
EST analysis
By using the UCSC genome browser (Kent et al. 2002), we located the relevant SNPs by entering the rs id numbers provided by dbSNP into the "position/search" input area. We used visual inspection to determine the nature of the alternative splicing event and tried to identify the allele that might be responsible for this splicing alteration. The browser was configured to display the "human mRNAs" and "spliced ESTs" tracks in "full" mode. Determining which allele was used in each EST was simplified by setting the "color track by codons or bases" option to "different mRNA bases" (follow the link on the track name to turn on this feature).

Acknowledgments
Jeremiah Faith and Susan Janicki gave insightful comments on the manuscript. The anonymous reviewers gave useful criticisms that helped improve the paper. A.J.O. and R.S. thank the DART Neurogenomics Alliance for support. X.R. and A.R.K. acknowledge support from NIH grant CA13106. A.R.R. acknowledges support from the Department of Biotechnology (India) grant BT/IN-BTOA/03/2005.

References

86 Genome Research
www.genome.org
spliceosome assembly but does not uniquely define the 5’ cleavage site. EMBO J. 183: 2533–2538.

Received July 3, 2007; accepted in revised form October 10, 2007.
Features of 5′-splice-site efficiency derived from disease-causing mutations and comparative genomics

Xavier Roca, Andrew J. Olson, Atmakuri R. Rao, et al.

Genome Res. 2008 18: 77-87 originally published online November 21, 2007
Access the most recent version at doi:10.1101/gr.6859308

Supplemental Material
http://genome.cshlp.org/content/suppl/2008/01/07/gr.6859308.DC1

References
This article cites 55 articles, 12 of which can be accessed free at:
http://genome.cshlp.org/content/18/1/77.full.html#ref-list-1

Open Access
Freely available online through the Genome Research Open Access option.

License
Freely available online through the Genome Research Open Access option.

Email Alerting Service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.